Neural Processing Letters (2006) 23:325–344
DOI 10.1007/s11063-006-9005-x

A New Cluster Validity for Data Clustering

XULEI YANG[1,⋆], QING SONG[1] and AIZE CAO[2]
[1]*Computer Control Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, s1-b4a-01, Singapore 639798, Singapore.*
e-mail: yangxulei@pmail.ntu.edu.sg [2]*School of Medicine, Vanderbilt University, Nahville, TN 37232, USA*

Abstract. Cluster validity has been widely used to evaluate the fitness of partitions produced by clustering algorithms. This paper presents a new validity, which is called the Vapnik–Chervonenkis-bound (VB) index, for data clustering. It is estimated based on the structural risk minimization (SRM) principle, which optimizes the bound simultaneously over both the distortion function (empirical risk) and the VC-dimension (model complexity). The smallest bound of the guaranteed risk achieved on some appropriate cluster number validates the best description of the data structure. We use the deterministic annealing (DA) algorithm as the underlying clustering technique to produce the partitions. Five numerical examples and two real data sets are used to illustrate the use of VB as a validity index. Its effectiveness is compared to several popular cluster-validity indexes. The results of comparative study show that the proposed VB index has high ability in producing a good cluster number estimate and in addition, it provides a new approach for cluster validity from the view of statistical learning theory.

Key words. cluster validity, data clustering, deterministic annealing, structural risk minimization, Vapnik–Chervonenkis-bound

1. Introduction

Clustering plays an important role in many engineering fields such as pattern recognition, system modelling, image processing, communication, data mining, and so on. The deterministic annealing (DA) algorithm, in particular, the mass-constrained DA algorithm, in which the annealing process with its phase transitions leads to a natural hierarchical clustering, is independent of the choice of the initial data configuration and has the ability to avoid poor local optima [19–21]. As reviewed in [21], the DA approach to clustering and its extensions has demonstrated substantial performance improvement over standard supervised and unsupervised learning methods. However, the DA clustering algorithm (i.e., the use of DA approach for data clustering) needs to pre-select the optimal cluster number of clusters, which is generally unknown in practical applications. Thus, an evaluation methodology is required to validate each of the partitions (according to $c = 2, 3, \ldots, c_{\max}$) and to obtain an optimal partition (or optimal number of clusters c^*). This quantitative evaluation is the subject of cluster validity. The mathematical

[⋆]Corresponding author.

formula used to compute the validation is referred to as a cluster validity index. In this paper, we use the DA algorithm as the underlying clustering technique to produce the partitions.

In the last three decades, many indexes have been proposed in the literature, which are used to measure the fitness of the partitions produced by clustering algorithms. Especially in fuzzy clustering, Bezdek first proposed two cluster validity indexes, the partition coefficient (PC) and partition entropy (PE) [1, 2]. Although these two indexes are widely cited in the literature, the major drawback is that they use only the fuzzy membership degrees for each cluster without considering the data structure of the clusters [11, 16]. To overcome this disadvantage, Xie and Beni [26] and Fukayama and Sugno [6] introduced new fuzzy validity criteria, denoted by XB and FS indexes, respectively, for evaluating fuzzy c-partitions by exploiting the concepts of compactness and separation. They combined, with a unique function, the properties of the fuzzy membership degrees and the structure of data, and therefore took into account the geometrical properties of the input data. Kwon extended the XB index to eliminate its tendency to monotonically decrease when the number of clusters approaches to the number of data points [13]. Most recently developed indexes focused on the modification or extension of the traditional compactness and separation to achieve better performance (e.g., [11, 15, 25, 27], etc.). Different from the indexes described above, which tend to focus on compactness and separation, the fuzzy hypervolume (FHV) and partition density (PD) validity functions proposed by Gath and Geva [7] have been developed based on measures of the degree of variance within each cluster. The variance measure was also used in some late proposed indexes such as in [4, 17].

In this paper, a new approach for clustering validity is proposed from the view of statistical learning theory. The proposed VC-bound (VB) index is estimated based on the structural risk minimization (SRM) principle [24], which optimizes the bound simultaneously over both the distortion function (empirical risk) and the VC-dimension (model complexity) to achieve the minimum of the guaranteed risk. The empirical risk is monotonically decreased with the increase of the cluster number c. While the model complexity is monotonically increased with the index of cluster number. The VB index aims to find a minimum value of the VC-bound (within the cluster number range) to make a trade-off between the minimizations of the empirical risk and model complexity. The corresponding cluster number and its partition validate the best description of the data structure. To construct the VB index, the empirical risk is represented by the distortion function of DA clustering algorithm, and the VC-dimension of system structure is found to be equal to the number of parameters of a set of specifically defined indicator functions. The validity indexes are considered to be independent of clustering algorithms. Most clustering algorithms can generate fuzzy partitions and cluster centers for a given data set. Although the proposed VB index is applicable to any fuzzy and crisp clustering algorithms, we only investigate its performance on the partitions produced by the DA clustering algorithm.

The rest of this paper is organized as follows. In Section 2, we review the DA clustering algorithm with several most cited validity indexes. The new cluster validity (i.e., VB index) is derived for data clustering in Section 3. The proposed index is estimated based on minimizing the bound over the guaranteed risk to reach the best description of the underlying data structure. In Section 4, the experimental results demonstrating the superiority of the VB index in appropriately determining the number of clusters, as compared to other well-known validity indexes, are provided for several numerical examples and real data sets. Conclusion is given in Section 5.

2. DA Clustering Algorithm and Several Popular Validity Indexes

2.1. DATA CLUSTERING BY DA ALGORITHM

Let the input data set be $X=\{x_1, x_2, \ldots, x_l\} \subset R^n$, where l is the number of input data points, n is the dimension of input space. Based on a measure of similarity (the most used one is the squared Euclidean distance), the data set is partitioned into c clusters whose centers are denoted by $V=\{v_1, v_2, \ldots, v_c\} \subset R^n$. Let $d(x_j, v_k)$ be the distance (squared Euclidean distance in [21]) between x_j and v_k, $p(v_k|x_j)$ be the association probability (membership) relating input point x_j with cluster center v_k, $p(x_j)$ be the source distribution. Then, the average expected distortion is given by

$$J_\mathrm{e}=\sum_{j=1}^{l}\sum_{k=1}^{c} p(x_j)p(v_k|x_j)d(x_j, v_k). \tag{1}$$

Minimization of J_e with respect to the free parameters $\{v_k, p(v_k|x_j\}$ would immediately produce a hard clustering solution, as it is always advantageous to fully assign an input point to the nearest cluster center [21]. To formulate the DA algorithm, Rose et al. [19] and Rose [20] recast the optimization problem as that of seeking the distribution, which minimizes J_e subject to a specified level of randomness. The level of randomness is, naturally, measured by the Shannon entropy as

$$H_\mathrm{s}=-\sum_{j=1}^{l} p(x_j)\sum_{k=1}^{c} p(v_k|x_j)\log p(v_k|x_j). \tag{2}$$

Then the DA algorithm is formulated as minimization of the Lagrangian

$$F=J_\mathrm{e}-\mathrm{TH}_\mathrm{s}. \tag{3}$$

It turns out [21] (according to the maximum entropy principle) that the resultant probability is the titled distribution and is given by

$$p(v_k|x_j)=\frac{p(v_k)\mathrm{e}^{-\frac{d(x_j, v_k)}{T}}}{Z_{x_j}}, \tag{4}$$

where $Z_{x_j} = \sum_{i=1}^{c} p(v_i)\mathrm{e}^{-\frac{d(x_j,v_i)}{T}}$ is the partition function, $p(v_k) = \sum_{j=1}^{l} p(x_j)p(v_k|x_j)$ is the mass probability of k(th) cluster, and T is the Lagrange multiplier, which bears an analogy to the temperature in statistical mechanics. Clearly, at limited high T, these are uniform distributions, each data point is equally associated with all clusters. These are extremely fuzzy associations. As T is lowered, the distributions become more discriminating and the associations less fuzzy. And at limited low T, the clustering becomes hard with each data point assigned to the nearest cluster with probability one.

To estimate the free parameter v_k, the effective cost to be minimized turns out to be the free energy (a well-know concept in statistical mechanics [19]) as follows.

$$F = \min_{\{p(v_k|x_j)\}} (J_\mathrm{e} - TH_\mathrm{s}) = -T\sum_{j=1}^{l} p(x_j)\log\sum_{i=1}^{c} p_i \mathrm{e}^{-\frac{d(x_j,v_i)}{T}}. \tag{5}$$

Based on (4), we can get the expression of cluster center v_k by minimizing (5), that is

$$v_k = \frac{\sum_{j=1}^{l} p(x_j)p(v_k|x_j)x_j}{\sum_{j=1}^{l} p(x_j)p(v_k|x_j)}. \tag{6}$$

Alternative updating (4) and (6) with phase transition gives the (mass constrained) DA algorithm, which has been shown superiority for data clustering and its extensions [21]. The pseudo-code of DA with the proposed validity index for data clustering is given in the next section.

During the annealing in T it is observed, that the cluster center remains at the mass center of the related cluster up to a critical value. At that point the representation undergoes a transition and the cluster center splits up in data space [9, 21]. The critical T_k^* for the phase transition for kth cluster can be approximately calculated as follows [21]

$$T_k^* = 2\lambda_{\max}(C_k(x)), \tag{7}$$

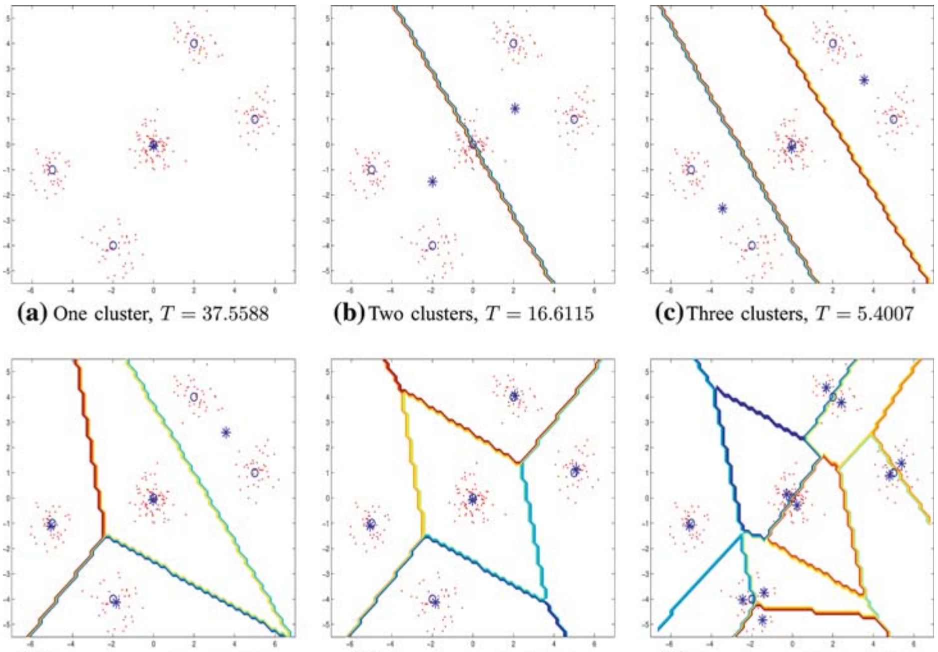
where $\lambda_{\max}(C_k(x))$ is the largest eigenvalue of the fuzzy covariance matrix of kth cluster, that is

$$C_k(x) = \sum_{j=1}^{l} p(x_j|v_k)(x_j - v_k)(x_j - v_k)^T \tag{8}$$

with the posteriori $p(x_j|v_k)$ obtained by the Bayes formula

$$p(x_j|v_k) = \frac{p(x_j)p(v_k|x_j)}{p(v_k)}. \tag{9}$$

Figure 1 shows the annealing process with its phase transitions on a simple example. The training set is generated from a mixture of five equal variance



(a) One cluster, $T = 37.5588$ **(b)** Two clusters, $T = 16.6115$ **(c)** Three clusters, $T = 5.4007$

(d) Four clusters, $T = 3.4204$ **(e)** Five clusters, T=2.1663 **(f)** Ten clusters, $T = 0.2595$

Figure 1. The annealing process with its phase transitions of DA clustering algorithm on a simple data. The original cluster centers are denoted by "o", the partitioned cluster centers are denoted by "*".

Gaussian whose centers are marked by "o". At high temperature, there is only one effective cluster represented by one cluster center, marked by "x", at the center of mass of the training set. As the temperature is lowered, the system undergoes phase transitions, which increase the number of effective clusters as shown in the figure.

2.2. SEVERAL POPULAR VALIDITY INDEXES

In this subsection, we brief review several popular fuzzy cluster validity indexes, which aims to: (1) provide a comparative study together with the proposed VB index and (2) investigate the performance of the existing fuzzy cluster validity indexes on the partitions produced by DA clustering algorithm. In the following the u_{kj} denotes the fuzzy membership of data point x_j belonging to the kth cluster.

Bezdek proposed two cluster validity indexes for fuzzy clustering, the PC and PE [1, 2], which were defined as

$$\mathrm{PC}(c) = \frac{1}{l} \sum_{j=1}^{l} \sum_{k=1}^{c} (u_{kj})^2 \qquad (10)$$

and

$$\mathrm{PE}(c) = -\frac{1}{l}\sum_{j=1}^{l}\sum_{k=1}^{c}(u_{kj})\log(u_{kj}). \tag{11}$$

The optimal partition (or an optimal value of c^*) is obtained by maximizing PC (or minimizing PE) with respect to $c = 2, 3, \ldots, c_{\max}$ because this provides compact clusters with higher values of u_{kj}.

The FHV and PD validity functions proposed by Gath and Geva [7] were defined by

$$\mathrm{FHV}(c) = \sum_{k=1}^{c} [\det(F_k)]^{1/2} \tag{12}$$

and

$$\mathrm{PD}(c) = \frac{\sum_{k=1}^{c} S_k}{\sum_{k=1}^{c} F_k}, \tag{13}$$

where the matrix F_k defined by

$$F_k = \frac{\sum_{j=1}^{l}(u_{kj})^m (x_j - v_k)(x_j - v_k)^T}{\sum_{j=1}^{l}(u_{kj})^m}, \tag{14}$$

denotes the fuzzy covariance matrix of cluster k; and the "sum of central members" S_k defined by

$$S_k = \sum_{j=1}^{l} u_{kj}, \quad \forall x_j \in \{x_j : (x_j - v_k)F_k^{-1}(x_j - v_k) < 1\} \tag{15}$$

takes into account only the points contributing to the core of the cluster, whose radii are the standard deviations of the cluster features. A fuzzy partition can be expected to have a low FHV (or high PD) value if the partition is tight. Thus, we find an optimal c^* by solving $\min_{2\le c\le c_{\max}} \mathrm{FHV}(c)$ (or $\max_{2\le c\le c_{\max}} \mathrm{PD}(c)$) to produce a best clustering performance for the data set X. Other indexes based on measures of the degree of variance within each cluster are [4, 17].

XB proposed a validity index [26] that focused on two properties: compactness and separation, defined as

$$\mathrm{XB}(c) = \frac{\sum_{j=1}^{l}\sum_{k=1}^{c}(u_{kj})^2\|x_j - v_k\|^2}{l\,\min_{i\neq k}\|v_i - v_k\|^2}. \tag{16}$$

The numerator indicates the compactness of the fuzzy partition, while the denominator indicates the strength of the separation between clusters. They stated that a good partition produces a small value for the compactness, and that well-separated cluster centers will produce a high value for the separation. Hence, the most desirable partition is obtained by minimizing XB for $c = 2, 3, \ldots, c_{\max}$.

FS also tried to model the cluster validation [6] by exploiting the concepts of compactness and separation, defined as

$$\mathrm{FS}(c) = J_{\mathrm{m}} - K_{\mathrm{m}} = \sum_{j=1}^{l}\sum_{k=1}^{c}(u_{kj})^m \|x_j - v_k\|^2 - \sum_{j=1}^{l}\sum_{k=1}^{c}(u_{kj})^m \|v_k - \bar{v}\|^2, \tag{17}$$

where $\bar{v}$ is the mean of the cluster centers, J_{m} is a compactness measure, and K_{m} measures the degree of separation between clusters. The optimal partition is obtained by minimizing FS with respect to $c = 2, 3, \ldots, c_{\max}$. Based on XB and FS indexes, lately developed indexes (e.g., [11, 13, 15, 25, 27]) focused on the modifying or extending of the traditional compactness and separation to achieve better performances.

3. The Proposed New Cluster Validity

In this section, we propose a novel practical cluster validity called VB index for data clustering from the view of statistical learning theory [24]. For this purpose, we need firstly to introduce the basics of SRM principle and related VC-bound as below.

3.1. SRM PRINCIPLE AND VC-BOUND

The principle of SRM is a key issue to obtain good generalization performances for a variety of learning machines, as e.g., the well-known support vector machines (SVMs). The SRM principle finds the function that for the fixed amount of data achieves the minimum of the guaranteed risk. To find the guaranteed risk, one has to use bounds, e.g., VC-bound, on the actual risk. Under the SRM, a set of admissible structures with the nested subsets is defined as follows

$$S_1 \subset S_2 \subset \cdots \subset S_c \ldots \tag{18}$$

with the non-decreasing VC-dimension of the structure

$$h_1 \leq h_2 \leq \cdots \leq h_c \ldots, \tag{19}$$

where $S_c = (Q^c(x_j, V) : V \in \Lambda_c), \forall j$ denotes the element of the structure at cluster number c, with a set of indicator functions $Q^c(x_j, V)$ of the empirical risk defined according to the problems under investigation [24]. The task of the SRM principle is to choose the element S_c of the structure for which the smallest bound on the real risk (the smallest guaranteed risk)[1]

$$\mathrm{VB} \leq R_{\mathrm{ems}} + \frac{\varepsilon}{2}\left(1 + \left(1 + R_{\mathrm{ems}}\frac{4}{\varepsilon}\right)\right)^{1/2} \tag{20}$$

[1] The bound (20) is based on the bias-variance dilemma [8]. The first term of inequality can be regarded as the bias (also viewed as the approximation error in [10, 22]) and the second term of the inequality can be regarded as the variance (also viewed as the estimation error in [10, 22]) from the view of

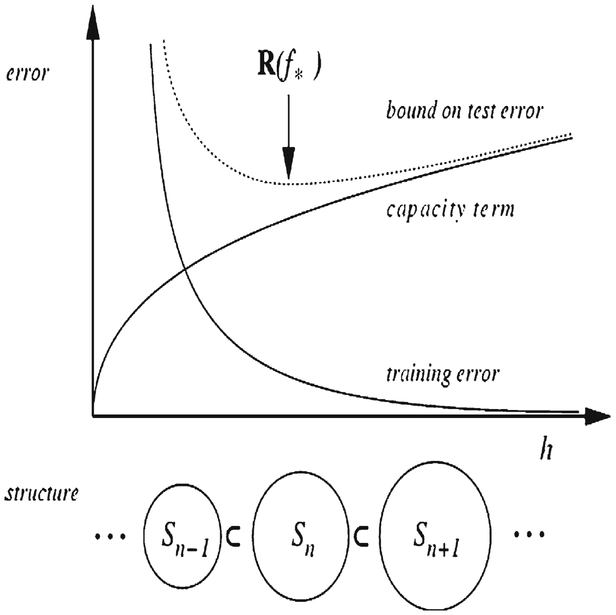


Figure 2. Graphical depiction of the SRM principle. The SRM picks a function f_*, which has small training error (empirical risk), and comes from an element of the structure that has low capacity h (confident interval), thus minimizing a risk bound in (20). The figure is taken from [22].

with

$$\varepsilon = \frac{h_c\left(\log\frac{2l}{h_c}+1\right)-\log\frac{\zeta}{4}}{l} \tag{21}$$

is achieved. Above $\zeta < 1$ is a constant (we set $\zeta = 0.01$ in all experiments). The first term of the right-hand side of the bound (20) represents the empirical risk and the second term is the confidence interval of the SRM based estimation. The bound on the risk is the sum of the empirical risk and of the confident interval. The empirical risk is decreased with the index of element of the structure (determined by cluster number in data clustering, refer to the next subsection), while the confidence interval is increased. The smallest bound of the risk is achieved on some appropriate element of the structure, as shown in Figure 2. We call the bound (20)

(*Footnote 1 continued*)
bias-variance dilemma. To achieve good overall performance, the bias and the variance of the learning would both have to be small. The VC dimension is the well-known technique to balance bias and variance raised in the literature of statistical learning theory. The important point here is that the definition can be used constructively to measure the capacity of a class of functions. The power of this approach stems from generic results about the rate of uniform convergence [8, 24]. The authors would like to thank the anonymous reviewer for pointing out the connection of the bound to the bias-variance dilemma.

above as "VC-bound[2]" (VC is the abbr. of Vapnik–Chervonenkis) in this paper. We apply it for data clustering to select the optimal cluster number in the next subsection. The induced cluster validity criteria is called VB index in this paper.

3.2. CLUSTER NUMBER SELECTION BY VB INDEX

Assume the given data is partitioned by the DA clustering algorithm, which results in certain partitions with according probabilities $p(v_k|x_j)$ $(k=1,2,\ldots,c$ and $j=1,2,\ldots,l)$ within the cluster number range $(c=2,3,\ldots,c_{\max})$. The key issue of applying (20) for cluster number selection is to construct a set of indicator functions specifically for DA clustering algorithm. Let's consider the function

$$P_k=\lim_{T\to 0} p(v_k|x_j)=\lim_{T\to 0}\frac{p(v_k)\mathrm{e}^{-d(x_j,v_k)/T}}{\sum_{i=1}^{c} p(v_i)\mathrm{e}^{-d(x_j,v_i)/T}}\approx\begin{cases}1, & \text{if } x_j\to v_k,\\ 0, & \text{otherwise,}\end{cases} \tag{22}$$

where $x_j\to v_k$ means the point j is assigned (clustered) into the cluster k at the limit low temperature $T\to 0$. Note the unique $P_k=1$ is reached if and only if the condition $d(x_j,v_k)<d(x_j,v_i)$ $(i\neq k; i,k\in[1,c])$ is satisfied. It bears the similarity to the step function. Since the independent variable $d(x_j,v_k)$ can be presented as an inner product of two (n)-dimensional vectors of the input space as

$$d(x_j,v_k)=\|x_j-v_k\|^2=\langle x_j-v_k\rangle\cdot\langle x_j-v_k\rangle=\left[(d_{kj}^1)^2+(d_{kj}^2)^2+\cdots+(d_{kj}^n)^2\right], \tag{23}$$

where d_{kj}^r $(r=1,2,\ldots,n)$ denotes the rth element of the vector x_j-v_k, and n is the dimensionality of the input data. Then the function (22) can be approximated by the step function as

$$P_k\approx\theta\left(\sum_{i=1}^{n}\gamma_{ki}\phi_{ki}(X^k)\right), \tag{24}$$

where γ_{ki} is the diameter parameter and ϕ_{ki} is linearly independent function with the subset $X^k\subset X$ of the input space.[3] We can now construct the indicator functions for the DA clustering algorithm as

$$Q^c(x_j,V)=\sum_{k=1}^{c}P_k\approx\sum_{k=1}^{c}\theta\left(\sum_{i=1}^{n}\gamma_{ki}\phi_{ki}(X^k)\right) \tag{25}$$

[2] We have to say that the name "VC-bound" is not popular for the bound (20) in the literature. We call it "VC-bound" here just because we used the VC-dimension in the bound, and in some sense, in honor of its originators.

[3] At the limit low temperature (hard clustering), we have $\sum_{i=1}^{n}\gamma_{ki}\phi_{ki}(X^k)=1$ for each data point with its nearest cluster k. This makes sure that $\phi_{ki}(X^k)$ is linearly independent without any assumption as in the hyperplane case, which needs one more bias parameter [24]. And v_k is a function of the subset $X^k\subset X$ of the input data points as derived in (6), in which the parameter γ_k has nonzero value only for $x_j\in X^k$ in the hard clustering partition.

It can be seen that the above constructed indicator functions, i.e., step functions (25), are linear in their parameters. According to [24] (chapter 4.11), the VC-dimension of a set of functions linear in their parameters is equal to the number of parameters, i.e.,

$$h_c = c \times n \tag{26}$$

for each nested subset S_c, such that the increase of cluster number is proportional to the increase of the estimated VC-dimension. To obtain good generalization performance one has to use the admissible structure (18) based on the set of indicator functions (25) to search for an optimal cluster number that minimizes the VC-bound (20).

Another important issue of applying (20) for cluster number selection is to represent the empirical risk in the VC-bound for DA clustering algorithm. It is well-known for the pattern recognition problem, the empirical risk in SRM estimation is defined based on the misclassified error as follows,

$$R_{\text{ems}} = \frac{1}{l}\sum_{j=1}^{l} \operatorname{sign}(y_j - \bar{y}_j), \tag{27}$$

where y_j is the actual class label of x_j and $\bar{y}_j$ is the estimated class label of x_j. Inspired by (27), a natural representation of empirical risk for clustering problem can be obtained based on the distortion measures. Hence, we define it as

$$R_{\text{ems}} = J_c / J_{\text{var}}, \tag{28}$$

where J_c is the distortion obtained by DA at cluster number c $(c = 2, 3, \ldots, c_{\max})$, which is defined by

$$J_c = \sum_{k=1}^{c}\sum_{j=1}^{l} p(x_j)p(v_k|x_j)d(x_j, v_k). \tag{29}$$

J_{var} is the variance of the input data, which is defined by

$$J_{\text{var}} = \frac{1}{l}\sum_{j=1}^{l} \|x_j - \bar{x}\|^2, \tag{30}$$

where $\bar{x} = \frac{1}{l}\sum_{j=1}^{l} x_j$ is the mean of the input data. The distortion J_c is scaled by the variance such that the induced R_{ems} is located inside the range of $[0, 1]$. As shown in the experimental results, this representation normally leads to the convex of the curve of VB index with the minimum value at the optimal cluster number. In conclusion, the proposed VB index is stated as: by evaluating the estimated VC-bound for each chosen cluster number by Equation (20), where h_c is defined by (26) and R_{ems} is defined by (28), we select the one that yields the minimum value of VB as the optimal cluster number.

3.3. PSEUDO-CODE OF THE PROPOSED METHOD

According to the above discussions, we now give a detailed pseudo-code of the DA clustering algorithm with the proposed VB index for cluster number selection as follows:

- Step (1) Set the maximum number of clusters $c_{\max}$[4], initial temperature $T_{\text{ini}} > 2T_1^*$ (see (7)), minimum temperature $T_{\min} = T_{\text{ini}}/1000$, convergence parameter $\varepsilon = 0.001$, cluster center $v_1 = \frac{1}{l}\sum_{j=1}^{l} x_j$ with mass probability $p(v_1) = 1$, source distribution $p(x_j) = \frac{1}{l}(j = 1, 2, \ldots, l)$, and $c = 1$.
- Step (2) Alternatively update the titled distribution (4), and the cluster centers (6) for $k = 1, 2, \ldots, c$ (fixed point iterations) until the maximum change in the cluster centers between consecutive iterations is less than the given threshold value ε.
- Step (3) Save the induced partition, and calculate the value of VB (20) (or any other cluster validity index introduced above) for the current cluster number c. If $c = c_{\max}$, set flag $= 0$ then go to step 5; otherwise, go to the next step.
- Step (4) If $T < T_{\min}$, set flag $= 1$ then go to step 5; Otherwise, let $T = \eta T$ $(0 < \eta < 1)$, and check condition of phase transition for $k = 1, 2, \ldots, c$, if critical T_k^* is reached for cluster k (see (7)), add a new cluster center by $v_{k+1} = v_k + \delta$ with $p(v_{k+1}) = p(v_k)/2$ and $p(v_k) = p(v_k)/2$, where δ is a small disturbance, and let $c \leftarrow c + 1$, then go to step 2.
- Step (5) If flag $= 0$, then select the optimal cluster number c^* by $c^* = \arg\min_c$ VB (here VB could be any other validity index), where $c = 2, 3, \ldots, c_{\max}$, and recover the saved partition according to the optimal cluster number c^*. If flag $= 1$, then indicate that the clustering procedure is failed.

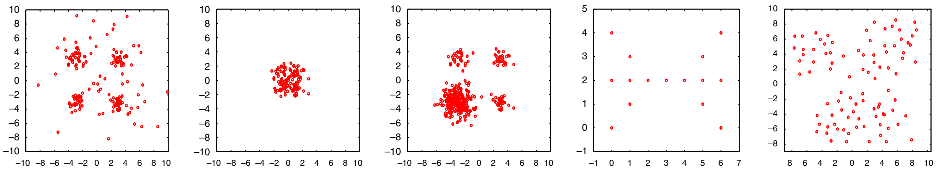
4. Experimental Results

To demonstrate the effectiveness and superiority of the proposed VB index for determining the optimal cluster number, we conducted extensive comparisons with other cluster validity indexes on five numerical examples and two real data sets. The proposed VB index was compared with six well-known fuzzy cluster validity indexes introduced in Section 2.2: Bezdek's PC and PE [1, 2], Xie and Beni's XB [26], Fukuyama and Sugenor's FS [6], and Gath and Geva's FHV and PD [7].

4.1. NUMERICAL EXAMPLES

For the purpose of data structure visualization, we only consider two-dimensional (2D) data sets in this subsection. We test the effectiveness and robustness of VB

[4] There is no general agreement on what value to use for $c_{\max}$, a rule of thumb that many investigators use is $c_{\max} \le (l)^{1/2}$ [16]. In this paper, we set $c_{\max} = 8$ in all experimental results.



(a) Noisy Data **(b)**Overlapped Data **(c)**Unbalanced Data **(d)**Buttery Data **(e)**Synthetic Data

Figure 3. The five data sets used in the numerical examples.

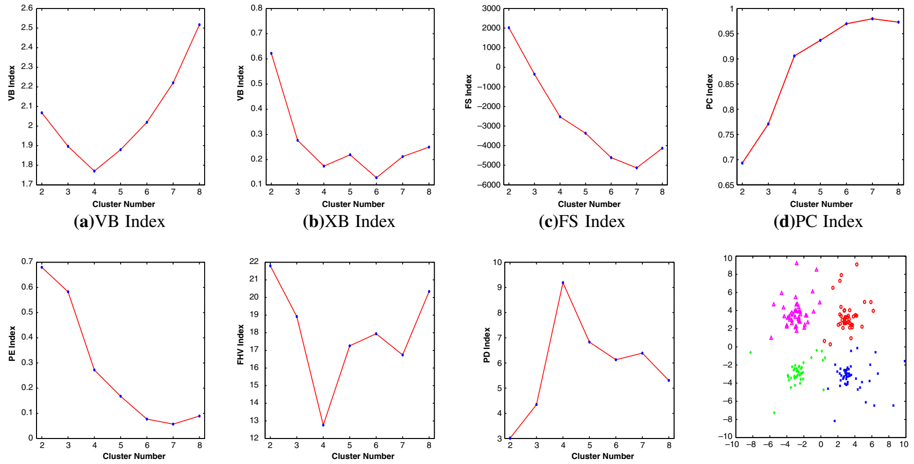
index for optimal cluster number selection on five specific data sets. The first three data tested are Noisy data shown in Figure 3(a), Overlapped data shown in Figure 3(b), and Unbalanced data shown in Figure 3(c). They are all derived from a simple data set, which is generated from four Gaussian distributions with means $[(-3,-3),(-3,3),(3,-3),(3,3)]$, variances $[0.6,0.6,0.6,0.6]$, and number $[30,30,30,30]$. The fourth data tested is Butterfly data [13] shown in Figure 3(d), in which 15 points have the positions of

$$\left\{ 0\ 0\ 0\ 1\ 1\ 1\ 2\ 3\ 4\ 5\ 5\ 5\ 6\ 6\ 6\quad 0\ 2\ 4\ 1\ 2\ 3\ 2\ 2\ 2\ 1\ 2\ 3\ 0\ 2\ 4 \right\}.$$

And the last one is Synthetic data where 90 data points are generated by random clicking as shown in Figure 3(e).

Example 1: Noisy Data: The Noisy data, as shown in Figure 3(a), is generated by contaminating the simple data set with 60 random noisy points. In practical applications, the presence of noise makes it harder to determine the number of clusters [5]. A good validity index should work well in the noisy situations. We plot the values of VB (20) with respect to the cluster number $c = 2, 3, \ldots, 8$ in Figure 4(a). It can be seen that the minimum of VB is reached at $c^* = 4$, so that the proposed VB index correctly reveals the underlying cluster number for this noisy data. We also plot the values of XB, FS, PC, PE, FHV, and PD with respect to cluster number c in Figure 4(b–g), respectively. Where only FHV and PD indexes can also find the correct cluster number $c^* = 4$ for this data set as shown in Figures 4(f) and (g). Among the unsuccessful indexes, PC, PE, and FS indexes indicate the optimal value of $c^* = 7$. XB index points to the optimal number $c^* = 6$; however, it indicates that $c^* = 4$ may be another good cluster number estimate as shown in Figure 4(b). For clustering result visualization, Figure 4(h) shows the partition obtained by DA clustering algorithm with $c^* = 4$.

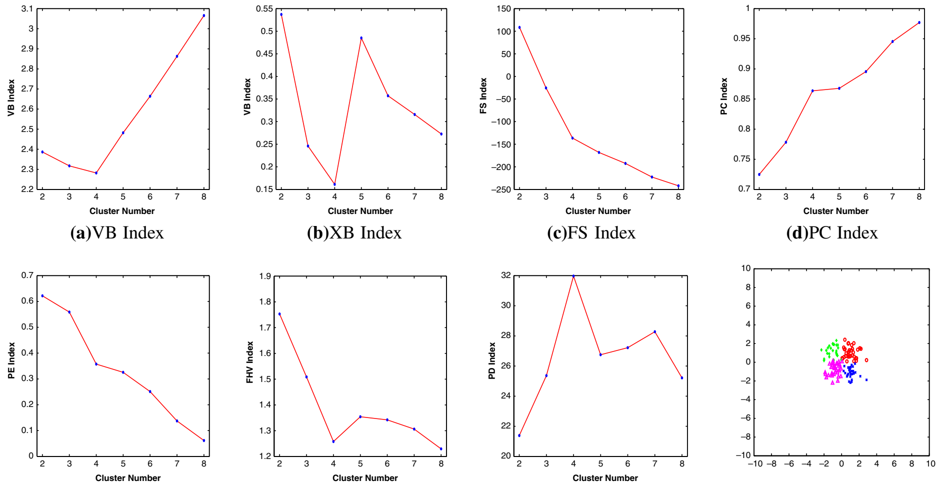
Example 2: Overlapped Data: The Overlapped data, as shown in Figure 3(b), is derived by moving the means of the four clusters in the simple data set from $[(-3,-3),(-3,3),(3,-3),(3,3)]$ to $[(-1,-1),(-1,1),(1,-1),(1,1)]$ but keeping other parameters unchanged. In practical applications, the data sets are normally overlapped. Hence, it is interesting to investigate the performances of existing validity indexes including the proposed VB on this overlapped data set. The plots of different validity indexes with different cluster number $c = 2, \ldots, 8$ are shown



(a)VB Index **(b)**XB Index **(c)**FS Index **(d)**PC Index

(e)PE Index **(f)**FHV Index **(g)**PD Index **(h)**Partition with c^* = 4

Figure 4. (a–g) Performances of different cluster indexes on Noisy data. (h) Clustering result of DA with $c^* = 4$ for Noisy data.



(a)VB Index **(b)**XB Index **(c)**FS Index **(d)**PC Index

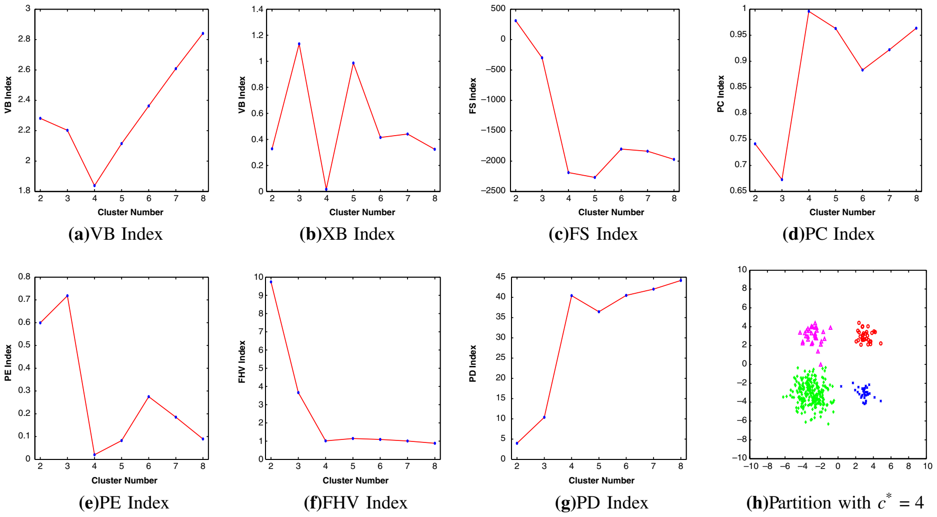
(e)PE Index **(f)**FHV Index **(g)**PD Index **(h)**Partition with c^* =4

Figure 5. (a–g) Performances of different cluster indexes on Overlapped data. (h) Clustering result of DA with $c^* = 4$ for Overlapped data.

in Figure 5. Four of them correctly point to the optimal cluster number $c^* = 4$ for this overlapped data set: VB, XB, and FHV indexes reach the minimum and PD gets the maximum at the point $c^* = 4$. The corresponding partition from DA clustering algorithm at $c^* = 4$ is shown in Figure 5(h), which correct reveals the underlying data structure. On the other hand, the unsuccessful indexes, PC, PE, and FS present a monotonic tendency of the cluster number c where they give a largest optimal cluster number estimate $c^* = 8$.

Example 3: Unbalanced Data: The Unbalanced data, as shown in Figure 3(c), is derived by changing the number of one cluster in the simple data set from 30 to 240 and the according variance from 0.6 to 1.2 but keeping other parameters unchanged. The situation that there are great difference in the number of samples in different clusters may occur in the practical applications, so it is also interesting to investigate this unbalanced data set. Figure 6 shows all the validity indexes with respect to the cluster number $c = 2, \ldots, 8$. It can be observed that the unbalance of the samples does not obviously affect the performances of validity indexes, the optimal number of clusters $c^* = 4$ for this unbalanced data set is correctly recognized by all validity indexes except PD and FS. The FS index incorrectly reveals the optimal number $c^* = 5$. Although PD points to $c^* = 6$ as the optimal value, it also indicates that $c^* = 4$ is a possible estimate. The partition of DA clustering algorithm with $c^* = 4$ is shown in Figure 6(h) for clustering result visualization. It provides a good description of the underlying data structure, only two points are misclassified.

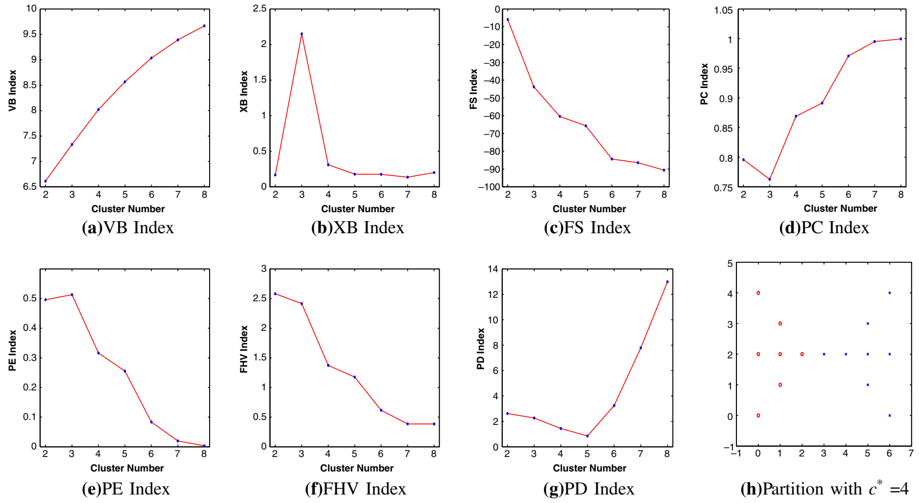
Example 4: Butterfly Data: The above examples considered three specific data sets with Gaussian distributions in a variety of situations. To show the generality of the proposed index, we consider two more data sets with nonGaussian distributions in this and the next examples. The Butterfly data, as shown in Figure 3(d), consists of 15 points with the preferred cluster number $c_{\text{opt}} = 2$. Figure 7 plots the performances of the seven cluster validity indexes on this data set with respect to the cluster number $c = 2, \ldots, 8$. It can be observed that only VB and XB indexes correctly reveal the optimal cluster number $c^* = 2$ as shown in Figures 7(a) and (b). PC, PE, and PD indicate that $c^* = 2$ is a possible optimal estimate though they point to a largest optimal cluster number $c^* = 8$. The other two indexes, i.e., FS



(a)VB Index **(b)**XB Index **(c)**FS Index **(d)**PC Index

(e)PE Index **(f)**FHV Index **(g)**PD Index **(h)**Partition with $c^* = 4$

Figure 6. (a–g) Performances of different cluster indexes on Unbalanced data. (h) Clustering result of DA with $c^* = 4$ for Unbalanced data.



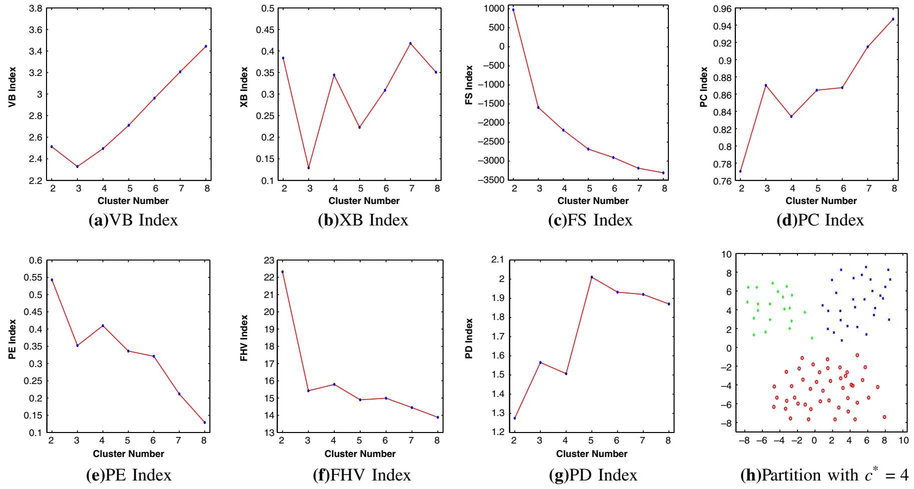
(**a**)VB Index (**b**)XB Index (**c**)FS Index (**d**)PC Index

(**e**)PE Index (**f**)FHV Index (**g**)PD Index (**h**)Partition with c^* =4

Figure 7. (a–g) Performances of different cluster indexes on Butterfly data. (h) Clustering result of DA with $c^* = 2$ for Butterfly data.

and FHV, cannot find the preferred cluster number: they reach the minimum at $c^* = 8$. The partitioning result of Butterfly data set by DA algorithm with $c^* = 2$ is shown in Figure 7(h) for visualization.

Example 5: Synthetic Data: The Synthetic data set, as shown in Figure 3(e), contains 90 randomly generated data points. These points are distributed in three assumed clusters with number 20, 30, and 40, respectively. The preferred cluster number for Synthetic data is $c_{\text{opt}} = 3$. The performances of different cluster validity indexes with different cluster number on this data set are shown in Figures 8(a–g).



(**a**)VB Index (**b**)XB Index (**c**)FS Index (**d**)PC Index

(**e**)PE Index (**f**)FHV Index (**g**)PD Index (**h**)Partition with c^* = 4

Figure 8. (a–g) Performances of different cluster indexes on Synthetic data. (h) Clustering result of DA with $c^* = 3$ for Synthetic data.

It can be seen that all the indexes except FS indicate that $c^* = 3$ is the optimal (VB and XB indexes) or a possible optimal (PC, PE, FHV, and PD indexes) cluster number. The unsuccessful index, i.e., FS, presents a monotonic tendency of the cluster number c where it gives a largest optimal cluster number estimate $c^* = 8$, as shown in Figure 8(c). The partitioning result of Synthetic data by DA algorithm with $c^* = 3$ is shown in Figure 8(h) for visualization.

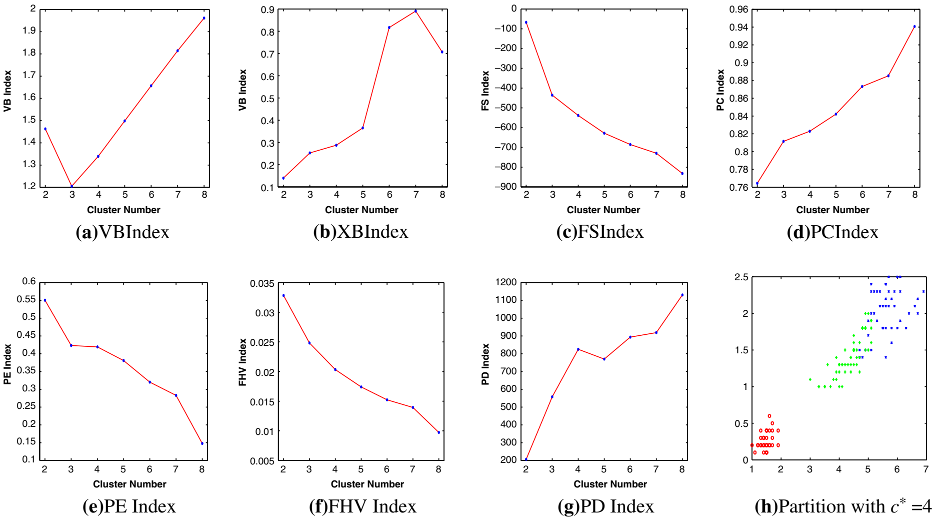
4.2. TWO REAL DATA SETS

The Iris and Wine data are investigated in this subsection for validity index comparisons. They are standard benchmarks in the machine learning literature and can be obtained from the UCI repository [23].

Example 6: Iris Data: The Iris data set contains three classes (Iris Setosa, Iris Versicolor and Iris Virginica) of 50 points each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are not linearly separable from each other. Thus, one can argue $c^* = 2$ or $c^* = 3$ for the Iris data set. The validity indexes of the Iris data set are shown in Figure 9. The proposed VB index reaches the minimum at $c^* = 3$ so that correctly reveals the underlying cluster number. The partition of DA clustering algorithm with $c^* = 3$ is shown in Figure 9(h) by Iris' two dominated features [i.e., petal length (PL) and petal width (PW)]. The XB index shows that $c^* = 2$, argued as another correct number for Iris [16], is the optimal cluster number estimate. The PD index regards $c^* = 8$ as the optimal number and indicates that $c^* = 4$ may be another optimal number. The results of other indexes (including PC, PE, FS, and FHV) are unexpected since they present a monotonic tendency of the cluster number c where they give a largest optimal cluster number estimate $c^* = 8$.

Example 7: Wine Data: As a final data set we present results from a wine recognition problem. The Wine data consists of 178 13-dimensional samples, which are a set of chemical analysis of three types of wine with number of samples 59, 71, and 48, respectively. Figure 10 shows the cluster validity results for the normalized Wine data set[5]. Two of them find the correct cluster number: both VB and XB indexes reach the minimum at the point of $c^* = 3$. The partition of DA clustering algorithm with $c^* = 3$ is shown in Figure 10(h) by Wine' first two principal components (taken from [18]). FHV indicates that $c^* = 3$ may be an optimal cluster number, though it reaches the minimum at $c^* = 8$. PC, PE, and FS indexes failed in this case, they show that $c^* = 7$ is the optimal cluster number. The PD index doesn't work in this case since it possesses a constant value with independence of cluster number c.

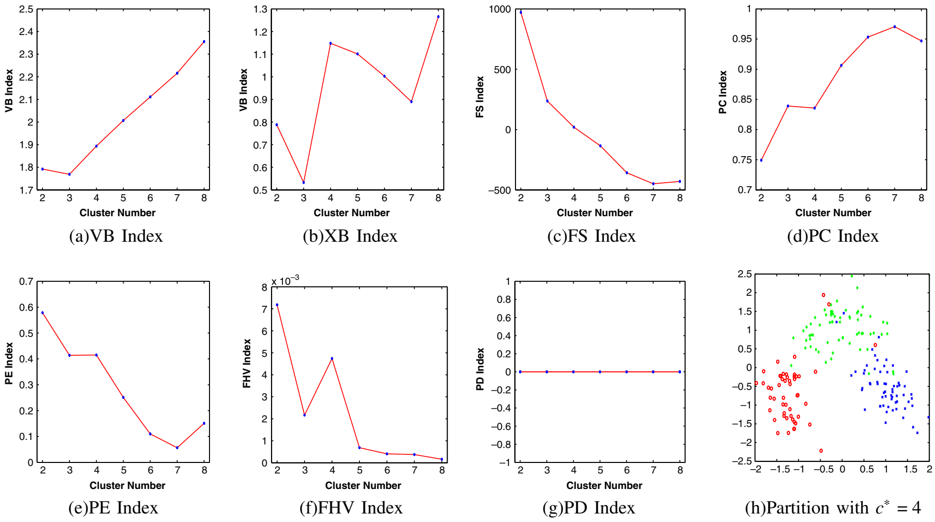
[5] Most clustering problems are solved by minimizing the constructed dispersion measures. In general, each dimension presents one characteristic of data in an n-dimensional data set where each characteristic has different dispersion. Thus, the results from minimizing the total dispersion measure may discard the effects of some characteristics, especially, for those that have small dispersion values. This situation frequently occurs in high-dimensional data sets. To use sufficiently all the information of characteristics, we shall normalize the data set [25].



(a)VBIndex **(b)**XBIndex **(c)**FSIndex **(d)**PCIndex

(e)PE Index **(f)**FHV Index **(g)**PD Index **(h)**Partition with $c^*=4$

Figure 9. (a–g) Performances of different cluster indexes on Iris data. (h) Clustering result of DA with $c^*=4$ for Iris data by its two dominated features (PL and PW).



(a)VB Index (b)XB Index (c)FS Index (d)PC Index

(e)PE Index (f)FHV Index (g)PD Index (h)Partition with $c^* = 4$

Figure 10. (a–g) Performances of different cluster indexes on Wine data. (h) Clustering result of DA with $c^*=4$ for Wine data by its first two principle components.

4.3. SUMMARY OF EXPERIMENTAL RESULTS

Table I summarizes the results obtained when the seven different validity indexes were applied to the seven different data sets tested. The column c_{opt} in Table I gives the optimal number of clusters for each data set, and the other columns show the optimal cluster numbers obtained using each validity index. The proposed VB index is the only index that correctly recognizes the number of clusters for all data

Table I. Values of c^* preferred by each cluster validity index for the seven data sets tested in this paper.

Data set	c_{opt}	VB	XB	FS	PC	PE	FHV	PD
Noisy	4	4	6(4)	7	7	7	4	4
Overlapped	4	4	4	8	8	8	4	4
Unbalanced	4	4	4	5	4	4	4	6
Butterfly	2	2	2	8	8(2)	8(2)	8	8(2)
Synthetic	3	3	3	8	8(3)	8(3)	8(3)	5(3)
Iris	2/3	3	2	8	8	8	8	8
Wine	3	3	3	7	7	7	8(3)	–

2/3 means both $c^* = 2$ and 3 are argued as optimal number for Iris; (x) means the number x inside the bracket is indicated as another good estimate; – means PD index doesnot work on Wine.

sets; hence VB is the most effective of the indexes considered. The XB index correctly identifies the optimal c^* in all data sets except the noisy data. It indicates that the correct number $c^* = 4$ may be another good cluster number estimate for the noisy data; hence XB is also an effective index. FHV and PD work well on most of the numerical examples with Gaussian distributions. However, they fail in most of the real data sets and numerical examples with nonGaussian distributions. Although they have been widely used for c-shells cluster number estimate, as in [12, 14], the practical applications of these two indexes need to be further investigated. On the other hand, PC, PE, and FS indexes present a monotonic tendency of the cluster number c on most of the data sets and therefore have the difficulty in revealing the correct cluster number. In particular, the FS indexes, failed in all data sets, hence it is proved as the most ineffective of the index considered. The PC and PE indexes find the correct cluster number for only the unbalanced data, hence they are also ineffective indexes. Note the direction of the monotonic tendency of PC (and PE) on fuzzy c-means FCM [3] clustering algorithm is the inverse of that on DA clustering algorithm. This is induced by the fact that they use only the membership from clustering algorithms. The fuzzy membership degree of FCM is increased with the increase of the cluster number assume that the fuzziness exponent keeps a constant, while for DA the probabilistic membership degree is decreased with the increase of the cluster number (as the temperature decreases from high value to low value). This observation indicates that some fuzzy cluster validity indexes are not suitable for evaluating the partitions produced by DA clustering algorithm.

5. Conclusion

Interest in clustering has increased recently because of new areas of application, such as data mining, image and speech processing, and bio-informatics. A central issue in these and other applications of clustering is how many clusters provide an appropriate description of the data. This is the main issue of cluster validity. In this paper, we have reviewed several validity indexes and then proposed a new validity index, which is called VB index, for data clustering. It is estimated based on the SRM principle in statistical learning theory, which optimizes the bound simultaneously over both the distortion function (empirical risk) and the VC-dimension (model complexity) to reach the minimum of the guaranteed risk, i.e., the best generalization performance. The smallest value of VB index on some appropriate cluster number validates the best description of the data structure. Its effectiveness and superiority are demonstrated by comparing to several popular cluster-validity indexes on five numerical examples and two real data sets. And the results of comparative study have shown that the proposed VB index has high ability in producing a good cluster number estimate.

Acknowledgment

The authors sincerely thank the anonymous reviewers for their insightful comments and valuable suggestions on earlier versions of this paper.

References

1. Bezdek, J. C.: Numerical taxonomy with fuzzy sets, *Journal of Mathematical Biology* **1**(1) (1974) 57–71.
2. Bezdek, J. C.: Cluster validity with fuzzy sets, *Journal of Cybernatics* **3**(3) (1974) 58–72.
3. Bezdek, J. C.: *Pattern Recognition with Fuzzy Objective Function ALgorithms*, Plenum Press, New York, 1981.
4. Boudraa, A. O.: Dynamic estimation of number of clusters in data sets, *Electronics Letters* **35**(19) (1999), 1606–1607.
5. Dave, R. N. and Krishnapuram, R.: Robust clustering methods: a unified view, *IEEE Transaction on Fuzzy System* **5**(2) (1997) 270–293.
6. Fukuyama, Y. and Sugeno, M.: A new method of choosing the number of clusters for the fuzzy c-means method, In: *Proceedings of the Fifth Fuzzy Systems Symposium*, pp. 247–250, 1989.
7. Gath, I. and Geva, A. B.: Unsupervised optimal fuzzy clustering, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **11**(7) (1989), 773–781.
8. Geman, S., Bienenstock, E. and Doursat R.: Neural Networks and the Bias/Variance Dilemma, *Neural Computation* **4**(1) (1992), 1–58.
9. Graepel, T., Burger, M. and Obermayer, K.: Self-organizing maps: generalizations and new optimization techniques, *Neurocomputing* **21**(1–3) (1998), 173–190.
10. Haykin, S.: *Neural Networks: A Comprehensive Foundation, 2nd edn.* Prentice Hall, NJ 1999.
11. Kim, D. W., Lee, K. H. and Lee, D.: On cluster validity index for estimation of the optimal number of fuzzy clusters, *Pattern Recognition* **37**(10) (2004), 2009–2025.
12. Krishnapuram, R., Nasraoui, O. and Frigui, H.: The fuzzy c-spherical shells algorithm: a new approach, *IEEE Transaction on Neural Networks* **3**(5) (1992), 663–671.
13. Kwon, S. H.: Cluster validity index for fuzzy clustering, *Electronics Letters* **34**(22) (1998), 2176–2177.
14. Man, Y. and Gath, I.: Detection and Separation of Ring-Shaped Clusters Using Fuzzy Clustering, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **16**(8) (1994), 855–861.

15. Pakhira, M. K., Bandyopadhyay, S. and Maulik, U.: Validity index for crisp and fuzzy clusters, *Pattern Recognition* **37**(3) (2004), 487–501.
16. Pal, N. R. and Bezdek, J. C.: On cluster validity for the fuzzy c-means model, *IEEE Transaction on Fuzzy Systems* **3**(3) (1995), 370–379.
17. Rezaee, M. R., Lelieveldt, B. P. F. and Reiber, J. H. C.: A new cluster validity index for the fuzzy c-mean, *Pattern Recognition Letters* **19**(3–4) (1998) 237–246.
18. Roberts, S. J., Holmes, C., Denison, D.: Minimum-entropy data partitioning using reversible jump Markov chain Monte Carlo, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **23**(8) (2001) 909–914.
19. Rose, K., Gurewitz, E. and Fox, G. C.: Statistical mechanics and phase transitions in clustering, *Physical Review letters* **65**(8) (1990) 945–948.
20. Rose, K., Gurewitz, E. and Fox, G. C.: Constrained clustering as an optimization method, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **15**(8) (1993), 785–794.
21. Rose, K.: Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, *Proceedings of the IEEE* **86**(11) (1998), 2210–2239.
22. Scholkopf, B. and Smola, A. J.: *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
23. *UCI Benchmark repository: A huge collection of artificial and real world data sets*, Availabe at http://www.ics.uci.edu/ mlearn.
24. Vapnik, V. N.: *Statistical Learning Theory*. Wiley Inc, New York, 1998.
25. Wu, K. L. and Yang, M. S.: A cluster validity index for fuzzy clustering, *Pattern Recognition Letters* **26**(9) (2005) 1275–1291.
26. Xie, X. L. and Beni, G.: A validity measure for fuzzy clustering, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **13**(8) (1991), 841–847.
27. Zahid, N., Limouri, M. and Essaid, A.: A new cluster-validity for fuzzy clustering, *Pattern Recognition* **32**(7) (1999) 1089–1097.